DOCUMENT RESUME

ED 382 687 TM 023 228

AUTHOR Motika, Robert T.; Chason, Walter M.

TITLE Performance of Angoff Model IV Linear Test Equating Using Total Test and Content Dimensional Sub-Test Designs in Small Groups of Examinees.

PUB DATE Apr 95

NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Education Majors; *Equated Scores; Estimation (Mathematics); French; Higher Education; Sampling; *Second Language Learning; Spanish; *Teacher Certification; Teacher Education; Test Construction; Test Content; *Test Format

IDENTIFIERS *Angoff Methods; *Linear Equating Method; Scale Drift

ABSTRACT

Test data from 200 examinees from the Spanish Teacher Certification Examination and 75 examinees from the French Teacher Certification Examination were used in a study of scale drift in sequentially equated test forms. Using sampling with replacement, 1,000 samples of 100 examinees each for Spanish and 1,000 samples of 50 each for French were created, and each sample was equated using Form A as the base form and equating through form B and form C back to form A. The Spanish forms were then broken down by item into three validated domains while the French forms were categorized into two domains. The forms were then equated by domain and the resulting equated domain scores summed to provide an estimate of total test performance. It was found that the whole test equating provided a better estimate based on scale drift than the summed scores of the equated domains. Nine tables. (Contains six references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ROBERT T. MOTIKA
WALTER M. CHASON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 382 687

Performance of Angoff Model IV Linear Test Equating Using Total Test and Content Dimensional Sub-Test Designs in Small Groups of Examinees.

Robert T. Motika and Walter M. Chason
Institute for Instructional Research and Practice
University of South Florida

Paper presented at the annual meeting of the
American Educational Research Association,
San Francisco, April 18, 1995

BEST COPY AVAILABLE

Performance of Angoff Model IV Linear Test Equating Using Total Test and Content Dimensional Sub-Test Designs in Small Groups of Examinees.

Robert T. Motika and Walter M. Chason
University of South Florida - Institute for Instructional Research and Practice

Abstract: Test data from 200 examinees from the Spanish Teacher Certification Examination and 75 examinees from the French Teacher Certification Examination were used in a study of scale drift in sequentially equated test forms. Using sampling with replacement, 1000 samples of 100 examinees each for Spanish and 1000 samples of 50 each for French were created and each sample was equated using form A as the base form and equating through form B and form C back to form A. The Spanish forms were then broken down by item into three validated domains while the French forms were categorized into two domains. The forms were then equated by domain and the resulting equated domain scores summed to provide an estimate of total test performance. It was found that the whole test equating provided a better estimate based on scale drift than the summed scores of the equated domains.

Purpose

The use of alternate test forms in certification and licensure testing programs is a commonly encountered method by which test security and fairness may be enhanced. The need for alternate forms and the equating procedures to link them when testing small groups of examinees presents problems for test administrators arising from the limited numbers of examinees and the imperfect understanding of how well these procedures relate to small groups of examinees. Linear equating using a common-item-non-equivalent groups design, a well-established procedure to relate raw scores on alternate test forms, was the equating method used in this study. This design has been referred to as the Angoff Design IV method of equating (Angoff, 1971). In this design, a new form is given to Group A, a previous form is given to Group B, and both forms include a shorter internal anchor test embedded within them. Responses to the internal anchor tests as well as the total raw response score are used to provide a conversion algorithm for comparison of the groups based on the following formula:

$$Y = AX + B = \frac{S_{yt} X}{S_{xt}} + \frac{\left(M_{yt} - S_{yt} M_{xt}\right)}{S_{xt}}$$

Where: S_{yt} = Standard deviation of test form Y for the total group, T

S_{xt} = Standard deviation of test form X for the total group, T

M_{yt} = Mean score of test form Y for the total group. T

M_{xt} = Mean score of test form X for the total group, T

This equating design offers several advantages to testing programs, including the use of non-random groups and the avoidance of the need to re-test persons. In spite of its widespread popularity, the performance of this linear equating technique when used with small groups of examinees has been examined by only a few researchers (e. g., Parshall et al., 1992).

An important aspect of linear equating using a common-item-nonequivalent-groups is the selection of items that will comprise the common or anchor test. Some studies have proposed that the use of content-representative anchors may improve the overall accuracy and precision of the equating (e. g., Klein & Jarjoura, 1985). The question of total-test and anchor-test dimensionality has also been examined previously, with similar findings. It has been concluded that unidimensionality of test and anchor items may improve the accuracy and precision of equated scores (Cook et al., 1983). In testing situations with limited numbers of examinees, the issue of dimensionality is made more troublesome since the small numbers make empirical examination of content dimensions, using a technique such as a principle components analysis nearly impossible.

The purpose of this research was to determine whether the accuracy and precision of the overall equating could be improved by dividing the existing test into smaller, independently equated, more content unidimensional subtests. Each resulting subtest would then be equated separately, summing the subscale equated scores to achieve a single equated measure of test performance. Equating test subscale scores in this manner may result in

enough overall improvement to warrant the additional equating procedures necessary, especially for situations involving small groups of examinees. This research examined the performance of Angoff Model IV linear equating for small samples of examinees taking Spanish and French foreign language teacher certification examinations. The research compared equating results when the entire test was equated using a single anchor (a whole test design), and when the equating was performed after the test and anchor are divided into smaller sub-scales based on item dimensionality (a sub-test design). The relative stability and precision of the equated scores resulting from these two methods were compared using the criteria of scale drift. This criteria is based on the results of equating through a chain or string of test forms. Scale drift is said to occur if the scores obtained from the direct equating of form A to form C is not the equivalent of equating form A to form C through an intermediate form B. In this study, a circular chain was used whereby the actual raw scores on form A were equated to form B, the equated B scores were equated to form C, and the equated C scores were then equated back to form A. After completing this circular chain of equatings, the raw scores on form A could be directly compared to the equated A scores resulting from the equating chain.

After equating each of these subtests separately, the overall test performance may be reconstituted by summing the scores obtained on each equated subtest portion. Due to the small sample sizes, identification of the items comprising the content dimensional sub-tests should be accomplished by means of expert examination of the items.

Method and Procedures

Data used in this study were gathered from Spanish and French foreign language teacher certification examinations. These examinations were administered in three different forms (A, B, and C) and the data set consisted of 200 first time test-takers per form for Spanish or 75 first time test takers per form for French. To assess the accuracy and precision of these equating procedures, an equating-chain format was used whereby form A was equated to form B, form B equated to form C, and form C back to form A. Using these linear equating algorithms, form A of each test could be equated through the chain back to itself (see Figure 1). This allows the comparison of the original raw score on A and the equated score on A calculated through the chain of A to B, B to C, and C to A. For the purposes of this study, 1000 samples of 100 examinees each for Spanish and 1000 samples of 50 each for French were randomly selected with replacement from the examinee pool of 200 and 75 respectively. The end result of this procedure was 1000 sets of equating algorithms which were then applied to the possible score range on form A to obtain 1000 values of the corresponding equated scores. The mean and standard deviation were then calculated for the 1000 values corresponding to each possible score point for form A. Scale drift was defined as the magnitude of any discrepancy between the initial score on form A and the equated score resulting from the three stage chain equating process.

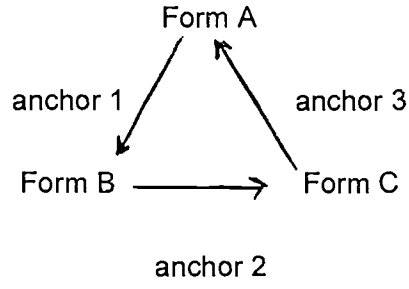

Figure 1. Equating Chain for Forms A, B, and C.

The amount of scale drift associated with equating when using the Angoff Model IV method was examined under two different scenarios: 1) with the entire test equated in one equating step using an anchor or common item subset comprising all common items on the two forms, and 2) with the test divided into sub-tests based on item dimensionality, each sub-test being equated independently of the other sub-tests, and the examinee's score being reconstituted at the final step of the process by summing each sub-test equated score.

Content experts in foreign language instruction were asked to group the items into similar categories or subscales based on the type of skill elicited from the examinee and to classify each item on the test into one of these resulting subscales. Examination of the items on the various forms by content matter experts yielded a three group typology for the Spanish test consisting of 1) a Receptive Skills sub-scale, 2) a Productive Skills sub-scale, and 3) a General/Pedagogical/Cultural sub-scale. Items comprising the Receptive Skills sub-scale were items designed to assess the examinee's ability to receive and understand information in the Spanish language. These items consisted of items involving knowledge of grammar and listening/reading comprehension. Productive skill items involved the production of language such as writing or speaking. The last category, General/Pedagogical and Cultural skills comprised items that measured the examinee's knowledge of Hispanic culture, teaching techniques and methods common in foreign language instruction, and other items not directly classifiable into the previous two categories.

For the French exam, the content experts divided the items into two sub-scales: 1) Grammar/Syntactical Items and 2) Cultural/Pedagogical Items. Items comprising the Grammar/Syntactical Items subscale consisted of such knowledge areas as grammar rules, rules of composition, and verb tense. The Cultural/Pedagogical subscale consisted of items assessing knowledge about French culture and social life, geography, and items addressing pedagogical issues in the teaching of foreign languages.

The compositions of the Spanish and French tests with respect to content representation are shown in Table 1 and Table 2 respectively. The data in these tables indicate that the equating anchors for both exams for the whole test equating closely mirrored the entire test in terms of percent composition among the different scales. The anchors for the subtest equating for each exam were comprised entirely of items within that particular scale.

Table 1

Percent Composition of Spanish Tests and Anchors

	Form A and Anchors				
	Form A	Anchor 1	Anchor 1-R	Anchor 1-P	Anchor 1-G
Receptive Skills	44%	49%	100%	0%	0%
Productive Skills	18%	20%	0%	100%	0%
General/ Skills	38%	32%	0%	0%	100%

	Form B and Anchors				
	Form B	Anchor 2	Anchor 2-R	Anchor 2-P	Anchor 2-G
Receptive Skills	45%	44%	100%	0%	0%
Productive Skills	19%	24%	0%	100%	0%
General/ Skills	36%	32%	0%	0%	100%

	Form C and Anchors				
	Form C	Anchor 3	Anchor 3-R	Anchor 3-P	Anchor 3-G
Receptive Skills	40%	41%	100%	0%	0%
Productive Skills	17%	18%	0%	100%	0%
General/ Skills	43%	41%	0%	0%	100%

Table 2

Percent Composition of French Tests and Anchors

	Form A and Anchors			
	Form A	Anchor 1	Anchor 1-G	Anchor 1-C
Grammar Skills	44%	49%	100%	0%
Cultural Skills	18%	20%	0%	100%

	Form B and Anchors			
	Form B	Anchor 2	Anchor 2-G	Anchor 2-C
Grammar Skills	45%	44%	100%	0%
Cultural Skills	19%	24%	0%	100%

	Form C and Anchors			
	Form C	Anchor 3	Anchor 3-G	Anchor 3-C
Grammar Skills	40%	41%	100%	0%
Cultural Skills	17%	18%	0%	100%

Results

Table 3 below shows the correlations between anchors and total test score for Spanish and French. It can be seen that for both tests the range of correlations were typical for tests of this nature (.78 - .98) with one correlation below .80, four correlations between .80 and .89, and seven correlations above .90. This is usually viewed as evidence for equivalent representation of the total test by the anchor.

Table 3

Spanish Whole Test Equating--Anchor and Rawscore Correlations

	Rawscore Form A	Rawscore Form B	Rawscore Form C
Anchor 1	.80	.89	--
Anchor 2	--	.96	.95
Anchor 3	.89	--	.94

French Whole Test Equating--Anchor and Rawscore Correlations

	Rawscore Form A	Rawscore Form B	Rawscore Form C
Anchor 1	.78	.81	--
Anchor 2	--	.94	.98
Anchor 3	.92	--	.97

Table 4 shows the correlations between subtest anchor scores and total subtest scores for Spanish and French. Given the small number of test items found in some of the subtests and their associated anchors, these correlations must be viewed with caution.

Table 4

Spanish SubTest Equating--Anchor and Rawscore Correlations

	Rawscore Form A			Rawscore Form B			Rawscore Form C		
	Subscale			Subscale			Subscale		
	R	P	G	R	P	G	R	P	G
Anchor 1	.93	.90	.87	.85	.83	.85	--	--	--
Anchor 2	--	--	--	.75	.77	.78	.76	.90	.80
Anchor 3	.90	.85	.88	--	--	--	.92	.89	.86

French SubTest Equating--Anchor and Rawscore Correlations

	Rawscore Form A		Rawscore Form B		Rawscore Form C	
	Subscale		Subscale		Subscale	
	G	C	G	C	G	C
Anchor 1	.69	.81	.70	.69	--	--
Anchor 2	--	--	.84	.94	.91	.72
Anchor 3	.86	.89	--	--	.65	.76

Figure 2 and Figure 3 show examinee performance across subscales. In these graphs, performance is indicated by the mean proportion correct for each subscale. In Figure 2, for example, it can be seen that examinee performance was highest for the items comprising the Receptive Skills subscale, next highest for the Productive Skills subscale, and

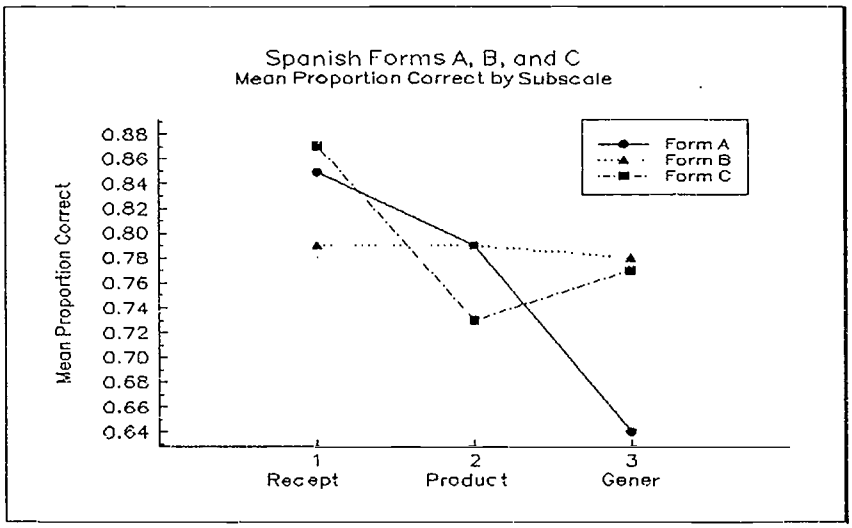

Figure 2

for the lowest items comprising the General\Cultural\Pedagogical subscale. This trend was evident for all three Spanish forms A, B and C.

Figure 3 shows that the performance of examinees on the French exam as measured by the mean proportion correct is highest for the Grammar/Syntactical subscale, lowest for the

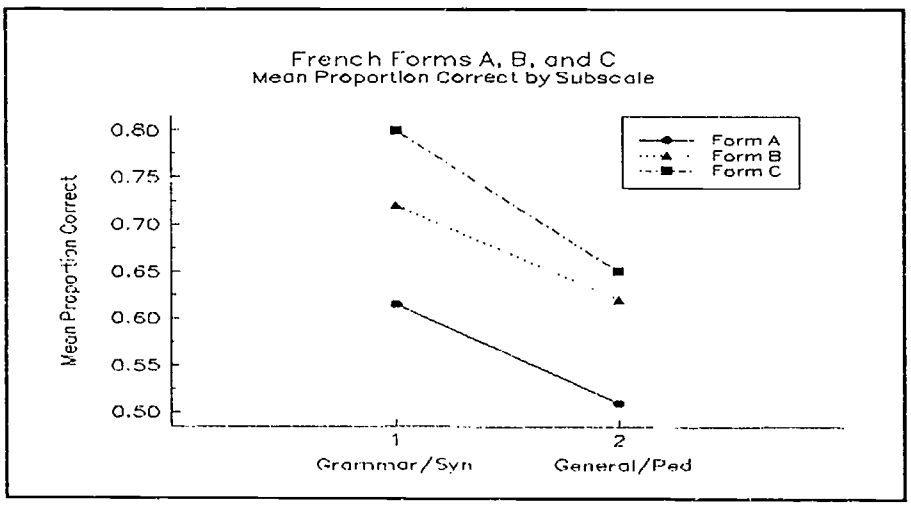

Figure 3

BEST COPY AVAILABLE

Cultural/Pedagogical subscale. This trend is also evident for all three forms of the French exam.

The measures of internal consistency seen in Table 5 indicate that the tests have a reasonable level of reliability based on this index.

Table 5

KR-20 Reliability Coefficients

	Form A	Form B	Form C
Spanish	.89	.88	.83
French	.92	.85	.83

The univariate statistics for each form, its anchors, and the subtests and their anchors are found in Tables 6 and 7.

Table 6

Univariate Statistics for Spanish Whole Test Equating Samples

	Form A	Anchor 1	Anchor 3	Form B	Anchor 1	Anchor 2	Form C	Anchor 2	Anchor 3
n	200	200	200	200	200	200	200	200	200
Item Number	87	41	39	84	41	39	89	39	39
Mean Raw Score	69.04	33.11	31.21	66.03	32.01	20.30	76.12	20.35	30.83
Raw Score Standard Deviation	10.20	5.29	4.40	10.07	4.99	3.23	8.69	2.83	4.03
Mean Pt. Biserial Correlation	.29	.27	.25	.26	.28	.22	.21	.20	.19

Table 7

Univariate Statistics for Spanish Sub-Test Equating Samples

Form A

	Subscale R	Anchor 1-R	Anchor 3-R	Subscale P	Anchor 1-P	Anchor 3-P	Subscale G	Anchor 1-G	Anchor 3-G
n	200	200	200	200	200	200	200	200	200
Item Number	40	20	16	16	8	7	35	13	16
Mean Raw Score	34.0	16.98	14.27	12.65	6.44	5.61	22.40	9.67	11.29
Raw Score Standard Deviation	4.43	2.68	1.90	2.62	1.53	1.40	4.87	2.21	2.39
Mean Pt. Biserial Correlation	.28	.28	.28	.32	.36	.33	.27	.26	.23

Form B

	Subscale R	Anchor 1-R	Anchor 2-R	Subscale P	Anchor 1-P	Anchor 2-P	Subscale G	Anchor 1-G	Anchor 2-G
n	200	200	200	200	200	200	200	200	200
Item Number	38	20	11	16	3	6	30	13	8
Mean Raw Score	30.12	16.23	9.39	12.59	5.86	4.40	23.32	9.88	6.45
Raw Score Standard Deviation	4.92	2.65	1.37	2.76	1.83	1.41	4.19	2.14	1.55
Mean Pt. Biserial Correlation	.25	.24	.20	.32	.36	.37	.24	.20	.26

Form C

	Subscale R	Anchor 2-R	Anchor 3-R	Subscale P	Anchor 2-P	Anchor 3-P	Subscale G	Anchor 2-G	Anchor 3-G
n	200	200	200	200	200	200	200	200	200
Item Number	36	11	16	15	6	7	38	8	16
Mean Raw Score	31.22	9.39	13.87	10.89	4.39	4.96	29.20	6.54	11.96
Raw Score Standard Deviation	3.68	1.42	1.82	2.39	1.21	1.55	4.84	1.40	2.15
Mean Pt. Biserial Correlation	.20	.19	.18	.21	.20	.25	.21	.23	.18

The regression of the Spanish equated Form A scores on the actual Spanish Form A scores for the whole test equating procedure yielded a slope of 0.913 with an intercept of 5.214. The slope and intercept of this regression, in the absence of any error of equating, would have been 1.0 and 0.0, respectively. The root-mean-squared-error (RMSE) of equating through the chain can be estimated using

$$\text{RMSE} = \sqrt{\frac{\sum_i n_i \left(X_i - X_i'\right)^2}{\sum_i n_i}}$$

where X_i is the i-th raw score on the given test date, n_i is the number of people obtaining raw score I on the given test date, and X_i' is the estimated equivalent of Xi estimated through the equating chain. The root-mean-squared-error (RMSE) for this equating was 1.23. The mean equating error or bias, which contributes to the RMSE, was calculated to be 0.987 using the following formula

$$\text{BIAS} = \overline{X} - \overline{X'}$$

where $\overline{X}$ is the mean of the raw scores and $\overline{X}'$ is the mean of the estimated equivalents of the raw scores.

The regression of Spanish Equated Form A scores on actual Spanish Form A scores for the sub-test equating procedure yielded a slope of 0.971 with an intercept of 0.439. Again, in the absence of equating error, these regression parameters should have been 1 and 0 respectively. The root-mean-squared-error (RMSE) for the sub-test equating was 1.62 and the BIAS was 1.596.

The regression of French Equated A scores on actual French Form A scores for the whole test equating procedure yielded a slope of 0.827 and an intercept of 8.43. The RMSE for this equating was found to be 2.18. The BIAS was found to be 1.70. The regression of French Equated A scores on actual French Form A scores for the subtest equating procedure

yielded a slope of 1.02 and an intercept of 1.138. The RMSE for this equating was found to be 2.64. The BIAS was found to be -2.60.

Since the score obtained on the multiple choice portion of the certification exam is combined with performance ratings of language ability, no cut score for these exams can be calculated directly. However, assuming average performance by the examinee of the performance section of the test, the critical area or likely range of the cut score on the Spanish examination is found in the raw score range (Form A) of about 60 to 69. The corresponding critical raw score range (or likely area of the cut score) for the French exam is about 46 to 55. Both raw score ranges as well as the corresponding equated scores from both the whole test and sub-test equating procedures are shown in Table 8.

Table 8 - Form A Raw Scores and Equated Scores for Spanish and French Exams

Spanish

Raw Score on Form A	Whole Test Equated Score	Mean Sub-Test Equated Score	Whole Test Equated Score Residual	SubTest Equated Score Residual
60	60.00	58.76	0.00	1.24
61	60.91	59.66	0.09	1.34
62	61.83	60.82	0.17	1.18
63	62.74	61.67	0.26	1.33
64	63.65	62.58	0.35	1.42
65	64.57	63.44	0.43	1.56
66	65.48	64.68	0.52	1.32
67	66.39	65.46	0.61	1.54
68	67.31	66.61	0.69	1.39
69	68.22	67.57	0.78	1.43

French

Raw Score on Form A	Whole Test Equated Score	Mean Sub-Test Equated Score	Whole Test Equated Score Residual	SubTest Equated Score Residual
46	46.46	48.01	-0.46	-2.01
47	47.29	48.99	-0.29	-1.99
48	48.11	50.09	-0.11	-2.09
49	48.94	51.13	0.06	-2.13
50	49.77	51.98	0.23	-1.98
51	50.59	53.06	0.41	-2.06
52	51.42	53.99	0.58	-1.99
53	52.25	55.04	0.75	-2.04
54	53.07	56.08	0.93	-2.08
55	53.90	57.05	1.10	-2.05

Discussion

Based on the magnitude of the RMSE and BIAS statistics for the whole test and subtests equating data, it can be concluded that there is evidence that the whole test method of equating is a better estimate of the equated score than the combination of the subtest equated scores. The RMSE for the Spanish form A equated to the "new" form A is 1.23 and a bias of

0.987. The RMSE for the Spanish subtest equating procedure was 1.62 and a bias of 1.596. Of the possible sources of equating error discussed by Kolen (1988), systematic error would be more likely to affect these data due to the non-equivalent groups design utilized. The lack of variable uniformity between the three forms of both the Spanish and French examinations makes determination of equivalent statistical groups difficult however it is suspected that the distribution shape differs markedly between the test form and its associated anchor. The skewness for form A is -1.25 while the skewness for its associated anchors is -1.00 and -1.12. The kurtosis for form A is 2.31 while the kurtosis for the associated anchors is 1.62 and 1.66. Table 9 shows that the distributions for the forms and anchors were not always equivalent. This is a likely source of systematic error.

Table 9

Form and Anchor Skewness and Kurtosis

Form	Skewness	Kurtosis
A	-1.25	2.31
Anchor 3	-1.00	1.66
Anchor 1	-1.12	1.62
B	-0.688	1.684
Anchor 1	-0.638	0.585
Anchor 2	-0.961	1.650
C	-0.393	-0.170
Anchor 3	-0.241	-0.399
Anchor 2	-0.274	-0.656

Conclusion

Our data indicate that whole test equating provides more accurate and precise results than does equating using subtest designs. However, results we obtained are sample dependent and may or may not be replicated with other data sets. Possible sources of error inherent in our sample could be the variability of number of anchor items by form and by

subtest, varied test length and consequently varied length of the subtests, and possible differences in examinee aptitude between forms due to the month of the year of administration of each form.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 508-600). Washington, DC: American Council of Education.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Peterson, N. S. (1983, April). An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Klein, L. W., & Jarjoura, D. (1987). The importance of content representation for common-item equating with non random groups. Journal of educational Measurement, 22(3), 197-206.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-items non-equivalent populations design. Applied Psychological Measurement, 11(3), 263-277.

Parshall, C. G., DuBose, P., & Kromrey, J. D. (1992, April). Common item linear equating in small samples of examinees: An empirical comparison of sample size effect. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 221-262). New York: National Council on Measurement in Education and American Council of Education.